Quantification of Prior Information Revised

Miroslav Kárný, Natalia Khailova, Petr Nedoma, Josef Böhm

Department of Adaptive Systems
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 4, 182 08 Prague 8, POB 18, Czech Republic
school@utia.cas.cz, tel. +420-02-6883421, fax +420-02-6884903

Prague, April 6, 2000

1 Introduction

Parameter estimation and output prediction [1, 2, 3] are well elaborated basic building blocks of many decision-making or control tasks. Their design, analysis and algorithms are permanently improved. This paper contributes to this development by proposing an improved exploitation of the prior information.

Anybody who deals with real-life decision problems appreciates the role of prior information. Within the context of Bayesian decision-making we consider [3, 4], data are combined with the information is built in

- a parameterised model $f(D|\Theta)$, i.e. the probability density function (pdf) of the observable data $D \in D^*$ conditioned on unknown parameters $\Theta \in \Theta^* \equiv$ a set of parameter values considered;
- a prior pdf $f(\Theta|K(\mathring{k}))$ where $K(\mathring{k}) \equiv (K_1, \ldots, K_{\mathring{k}})$ with $K_k \in K^* \equiv \{K_1, \ldots, K_{\mathring{k}}\} \equiv$ a collection of individual pieces of prior information K_k. The cardinality $\mathring{k}$ of the set K^* is always finite.

(The symbols X^*, $\mathring{X}$ mean a set of X values and its cardinality, respectively.)

Here, we deal with a construction of prior pdf that complements former procedures [5] that quantify prior information, i.e. construct $f(\Theta|K(\mathring{k}))$.

Information elicitation lies in roots of Bayesian paradigm and it is performed by a range of techniques, for a review, see e.g. [4]. Often, the information elicitation is based on a controlled dialog of "Bayesian expert" (statistician, evaluator) with expert providing the information (provider). This dialog allows the statistician to make a thorough information elicitation and to take into account the expert's competency. This way, however, requires high skill of both parties and addresses mostly low dimensional cases accessible to restricted abilities of human beings. In higher dimensions, the functional form of the prior pdf is mostly fixed. Such pdfs are called hyper-parameterised priors. Hyper-parameters of such a pdf are estimated/eliminated by maximum likelihood (second level maximum likelihood) or by assigning them prior pdf (hierarchical priors) [4].

Need for a further research has arisen in connection with a project of computerised off line design of adaptive controllers [6]. This project creates an algorithmic system that converts user's objectives, restrictions, prior information and available data into fully pre-tuned adaptive controller. In this context, processing of the prior information has specificity that are not covered by ready solutions [4]. Let us name the most important of them:

- prediction of the closed-loop (sic) behaviour is done using the posterior pdf of parameters: it relies to an extreme extent on quality of the information processing used;
- information provider and evaluator is usually single expert with other than statistical background: he can assign himself neither credibility level nor prevent partial contradictions and dependencies among individual pieces of prior information;

- list of usually available (global) information items can be gradually built (information on static gain, samples from a frequency response etc.);
- data obtained from a complex simulation model (different from that identified) are often available as an important information source.

The technique described in this paper covers these aspects in a wider extent than the techniques ready. It can be seen as elicitation with hyper-parameterised priors. It fits well to important cases with naturally conjugate priors [4] and it is able to cope with high-dimensional situations. These features are paid by a restricted expressive ability of considered priors. For instance, the number of modes of such a prior can be hardly influenced. Its change is tightly related to a change of the parameterised model. This is, however, topic for other paper.

The advocated algorithm consists of the following steps:

1. the individual information pieces are converted into a common form called fictitious data;
2. posterior degrees of belief (conditioned by available real data) are assigned to individual information pieces;
3. the constructed posterior pdf on estimated parameters is taken as weighted geometric mean of posterior pdfs obtained for individual prior information items; the degrees of belief found in the step 2 are used as weights.

The described technique is successfully implemented in a running version of the system DESIGNER [6, 7] and serves well in it. We believe that it is applicable in a wider range of applications. For this reason, explanations are kept at relatively general level. We rely on a classical reference [3] but we have tried to make the paper relatively self-containing.

2 Preliminaries and addressed problem

Let us assume that the available data D consist of a sequence of output – input pairs measured on the system of interest. The sequence $D \equiv d(\mathring{t}) \equiv (d_1, \ldots, d_t, \ldots, d_{\mathring{t}})$ is considered up to a horizon $\mathring{t} < \infty$ that determines cardinality of the time-indices set $t^* = \{1, \ldots, \mathring{t}\}$. The individual data items d_t can be split into $d_t \equiv (y_t, u_t) \equiv$ (output, input) of the inspected system. The joint pdf $f(D|\Theta)$ of data D conditioned by unknown parameters Θ is assumed to be of the following (relatively general) form

$$\begin{aligned} f(D|\Theta) &\equiv f(d(\mathring{t})|\Theta) = \prod_{t \in t^*} f(y_t|u_t, d(t-1), \Theta) f(u_t|d(t-1), \Theta) = \qquad (1) \\ &= \prod_{t \in t^*} m([y_t, \psi_t']', \Theta) f(u_t|d(t-1)) \propto \prod_{t \in t^*} m(\Psi_t, \Theta) \equiv \mathcal{L}(\Theta, d(\mathring{t})). \end{aligned}$$

In (1), we applied the chain rule for pdfs, and used

- natural conditions of control [3], i.e. the assumption that the parameters are unknown to the (random) input generator $f(u_t|d(t-1), \Theta)$;
- the assumption that the conditional pdf $f(y_t|u_t, d(t-1), \Theta)$, describing reactions of the (system) output y_t on the past data $d(t-1)$ and the current (system) input u_t, is a time invariant function $m(\Psi_t, \Theta)$ of a finite-dimensional vector (called data vector) $\Psi'_t = [y_t, \psi'_t] \equiv$ [regressand, regression vector] and unknown parameter Θ.

  The vector ψ_t is the value of a finite dimensional function, say Φ, that can be updated recursively using just its previous value and the newest data u_t, y_{t-1}, i.e. $\psi_t = \Phi(u_t, y_{t-1}, \psi_{t-1})$;
- the pdfs $m(\Psi_t, \Theta)$, $t \in t^*$ that are called parameterised model (of the system);
- the sign $'$ that denotes transposition,
- the symbol $\propto$ that means equality up to a normalising factor,
- the symbol $\mathcal{L}(\Theta, d(t))$, which is reserved for likelihood function evaluated for data $d(t)$, i.e. the product of parameterised models of the system.

Let the available information be $K(\mathring{k}) \equiv (K_1, \ldots, K_k, \ldots, K_{\mathring{k}})$. The individual pieces K_k can be of a various nature with a varying degree of precision and compatibility. For this reason, we express them in a common form called *fictitious data*. The data vector Ψ_k is called the fictitious data vector expressing the information item K_k (at "fictitious" time k) iff

$$f(\Theta|K_k) \propto m(\Psi_k, \Theta) p(\Theta) \tag{2}$$

where $p(\Theta)$ is a given pre-prior pdf that expresses a common (usually vague) information before incorporating any part of $K(\mathring{k})$. The definition (2) means that the individual information item K_k can be built-in into the prior pdf by using the standard Bayes rule [3] with the corresponding fictitious data vector Ψ_k inserted into the parameterised model $m(\cdot, \cdot)$ as a measured one.

At general level, pdfs $f(\Theta|\bullet)$ describe belief into "appropriateness" of parameters of the system model viewed as an input-output operator. Fictitious data express an information item that should be incorporated into $f(\Theta|\bullet)$ as a (thought) input-output behaviour of the system. They can be designed by obeying the operationally understandable command:

> Try to express an individual piece of information as the "expected" response of the considered system to a specific stimulus.

This fits well to experimental origin of such characteristics like static gain or frequency response. The way how to "translate" usual information items into fictitious data can be learnt through the guiding examples given in Section 6.2.

The notion of fictitious data is auxiliary one but we found it operationally very useful. It allows us:

- to combine parametric and data-based information (we consider both historical data and data obtained by simulation),
- to elicitate the information from practitioners without bothering them with detailed description of model,
- to decrease dimensionality of the space to be considered by human being (he can describe individual information pieces separately and, moreover, dimensionality of the space of data vectors Ψ_t is mostly lower than the dimensionality of space of the corresponding sufficient statistics),
- to make uncertainties of individual information pieces comparable by normalising fictitious data so that their uncertainty resembles expected uncertainty in real data.

Combination of mutually dependent information items is the crucial problem to be addressed irrespectively of the way they are expressed. It is hard problem, which has no deductive solution unless we are able or willing to specify dependencies among individual information items.

We present a solution that rely on use of fictitious data and availability of a set of real data. The joint use of these information sources complements and improves our former combination techniques [8, 9].

3 General solution

We have:

- the parameterised model of the system, see (1), $f(y_t|u_t, d(t-1), \Theta) = m(\Psi_t, \Theta)$,
- the available information $K(\mathring{k})$ on unknown parameters, for which
$$f(y_t|u_t, d(t-1), K(\mathring{k}), \Theta) = f(y_t|u_t, d(t-1), \Theta) \equiv m(\Psi_t, \Theta), \qquad (3)$$
i.e. the processed prior information influences the system description through the prior pdf of parameters only,
- a collection of pdfs $f(\Theta|K_k) \propto m(\Psi_k, \Theta)p(\Theta)$, determined by the fictitious data vectors $\Psi_k,\ k \in k^*$ and by the pre-prior pdf $p(\Theta)$, see (2),
- a sequence of measured data $d(\mathring{t})$.

We search for approximation $\hat{f}(\Theta|d(\mathring{t}), K(\mathring{k}))$ of the posterior pdf $f(\Theta|d(\mathring{t}), K(\mathring{k}))$ without explicitly specifying mutual dependencies of information items.

The posterior pdfs that exploit just individual information items are given by the Bayes rule [3]. Use of equations (1), (3), (2) leads to

$$f(\Theta|d(\mathring{t}), K_k) \propto \mathcal{L}(\Theta, d(\mathring{t}))p(\Theta)m(\Psi_k, \Theta),\ k \in k^*. \qquad (4)$$

Let $\alpha(K_k|d(\mathring{t}))$ denote degree of belief that the predictive pdf $f(d(\mathring{t})|K_k)$ – computed on the basis data $d(\mathring{t})$ and kth information item only – has the smallest Kullback-Leibler distance [10] to the unknown "objective" pdf of data $f(d(\mathring{t})$. It can be shown [11] that the Bayes rule provides the needed $\alpha(K_k|d(\mathring{t}))$ by modifying their prior values $\alpha(K_k),\ k \in k^*$.

We use uniform prior probabilities (= degrees of belief) $\alpha(K_k),\ k \in k^*$ reflecting our lack of preferences among alternative pieces of information. Thus, under the natural conditions of control,

$$\alpha(K_k|d(\mathring{t})) \underbrace{\propto}_{Bayes\ rule} f(d(\mathring{t})|K_k)\frac{1}{\mathring{k}} \underbrace{\propto}_{marginalisation} \int f(d(\mathring{t}),\Theta|K_k)\,d\Theta$$
$$\underbrace{\propto}_{(1),(2)} \frac{\int \mathcal{L}(\Theta,d(\mathring{t}))p(\Theta)m(\Psi_k,\Theta)\,d\Theta}{\int p(\Theta)m(\Psi_k,\Theta)\,d\Theta} \tag{5}$$

(the symbol $\underbrace{\quad}_{reference}$ points to justification of the corresponding step).

It is reasonable to interpret $\alpha(K_k|d(\mathring{t})),\ k \in k^*$ as degrees of belief into the individual information pieces. Then, the geometric mean of pdfs (4)

$$\hat{f}(\Theta|d(\mathring{t}),K(\mathring{k})) \propto \prod_{k\in k^*}[f(\Theta|d(\mathring{t}),K_k)]^{\alpha(K_k|d(\mathring{t}))}$$

has the smallest *expected* Kullback-Leibler distance to the unknown $f(\Theta|d(\mathring{t}),K(\mathring{k}))$.

We assume that the discussed probabilities $\alpha(K_k|d(\mathring{t}))$ sum to unity. It means that we use all information pieces available and avoid the problem known among Bayesians as the "problem of something else" [3].

The form of the presented solution can be simplified further on. Let us introduce the posterior likelihood corresponding to the flat pre-prior pdf $p(\Theta)$

$$L(\Theta,d(\mathring{t})) \equiv \mathcal{L}(\Theta,d(\mathring{t}))p(\Theta),\ \text{ with } L(\Theta,d(0)) \equiv p(\Theta) \tag{6}$$

Then, equation (5) implies

$$\alpha(K_k|d(\mathring{t})) \propto \frac{\int L(\Theta,d(\mathring{t}))m(\Psi_k,\Theta)\,d\Theta}{\int L(\Theta,d(0))m(\Psi_k,\Theta)\,d\Theta} \equiv w_k \tag{7}$$

Using the definition (6) and taking into account that $\alpha(K_k|d(\mathring{t}))$ sum to unity, we can re-write the estimate of $f(\Theta|d(\mathring{t}),K(\mathring{k}))$ into the form

$$\hat{f}(\Theta|d(\mathring{t}),K(\mathring{k})) \propto L(\Theta,d(\mathring{t}))\prod_{k\in k^*}[m(\Psi_k,\Theta)]^{\alpha(K_k|d(\mathring{t}))}.$$

Remarks

1. The proposed combination of information pieces is a conservative one: each information item gets a weight $\alpha < 1$. We found it practically advantageous as information providers are inclined to repeat a single piece

of information many times (often in different forms). The probabilistic weighting guarantees that the information piece is incorporated just once. Generally, use of such a weighting avoids dangerous over-fitting of information pieces that prevents efficient use of real data.

2. The obtained result is independent of the order of the processing of prior information unlike its predecessors. Also, computational complexity of this type of processing is relatively small for usual data sets available for off-line processing.

3. The need for a recursive, order-dependent, approximation (similar to [5]) arises if the cardinality of the set of prior information items is too large. It happens, typically, when data taken from a complex realistic simulation model is used. A feasible order-independent solution applicable whenever the data set reflects internally homogeneous information pieces is presented in Section 7.

4 Solution for exponential family

Let us specialise the general result of Section 3 to the exponential family [12] and naturally conjugate prior pdf [4]. This combination guarantees that the functional form of the prior pdf is preserved in the posterior pdf, too. This gave to naturally conjugate priors an alternative name self-reproducing priors. Moreover, data determine this functional form through a finite-dimensional sufficient statistics. Consequently, the functional Bayesian rule converts into algebraic recursion for the corresponding sufficient statistics. This fact allows us to perform estimation on digital computers what explains importance of this case.

The parameterised model $m(\Psi_t, \Theta)$ belongs to exponential family iff it has the form

$$f(y_t|u_t, d(t-1), \Theta) \equiv m(\Psi_t, \Theta) \propto A(\Theta) \exp < B(\Theta), C(\Psi_t) > \tag{8}$$

with a scalar function $A(\cdot) \geq 0$ and multivariate functions $B(\cdot), C(\cdot)$ of respective arguments and compatible dimensions. The symbol $< \cdot, \cdot >$ denotes scalar product. Typically, $< B, C >= B'C$ for vectors B, C or $< B, C >= \mathrm{tr}[B'C]$ for matrices ($\mathrm{tr}[\cdot]$ denotes matrix trace).

The likelihood function related to this model has the simple form

$$\mathcal{L}(\Theta, d(\mathring{t})) = A^{\nu_{\mathring{t}}}(\Theta) \exp < B(\Theta), V_{\mathring{t}} > . \tag{9}$$

The sufficient statistics $V_{\mathring{t}}, \nu_{\mathring{t}}$ evolve according to the following recursions

$$V_t = V_{t-1} + C(\Psi_t),\ \nu_t = \nu_{t-1} + 1,\ t \in t^*. \tag{10}$$

These recursions start from zero initial values $V_0 = 0,\ \nu_0 = 0$.

We select the pre-prior pdf $p(\Theta)$ in the naturally conjugate form. It formally coincides with the likelihood (9) for some statistics $\bar{V}, \bar{\nu}$. Generally, we should restrict the support of this likelihood to a more narrow set Θ^* than the domain of the likelihood. It is simply done by multiplying this pre-prior likelihood by the indicator $\chi(\Theta)$ of this set ($\chi = 1$ on Θ^* and zero outside of it).

With this choice, the posterior likelihood has formally the same form as the pre-prior one. Its statistics evolve according to the recursions (10) with initial conditions determined by the pre-prior pdf $p(\Theta)$

$$V_0 = \bar{V},\ \nu_0 = \bar{\nu}.$$

For the exponential family and reproducing priors, it is easy to verify that the posterior pdf expressing compromise among the individual pieces of the prior information preserve the same form. It is explicitly specified in next Section that summarises the overall algorithm for the exponential family.

5 Algorithm for exponential family

Preparatory phase

1. Select parameterised model $m(\Psi, \Theta) = A(\Theta) \exp < B(\Theta), C(\Psi) >$ (its form and structure).

2. Specify pre-prior information by selecting statistics $\bar{V}$, $\bar{\nu}$ (typically, corresponding to a very flat pdf) and by specifying the set Θ^* on which its support should be reduced.

3. Compute statistics corresponding to the posterior likelihood function assigned to the pre-prior pdf, i.e.

$$V_t = V_{t-1} + C(\Psi_t),\ \nu_t = \nu_{t-1} + 1,\ t \in t^*,\ V_0 = \bar{V},\ \nu_0 = \bar{\nu}.$$

   The final likelihood function corresponding to the vague pre-prior pdf $p(\Theta)$ has the form

$$L(\Theta, d(\mathring{t})) \equiv L(\Theta, V_{\mathring{t}}, \nu_{\mathring{t}}) = A^{\nu_{\mathring{t}}}(\Theta) \exp < B(\Theta), V_{\mathring{t}} > \chi(\Theta).$$

4. Prepare the procedure for integrating this likelihood function

$$\mathcal{I}(V, \nu) = \int A^{\nu}(\Theta) \exp < B(\Theta), V > \chi(\Theta)\, d\Theta. \tag{11}$$

Processing of information items

1. Set the auxiliary variable $norm = 0$ and

2. Do, for $k \in k^*$, while information pieces are available:

   (a) Express the information item K_k through fictitious data Ψ_k.

(b) Compute the weight w_k, see (7), and add it to $norm$.

$$w_k = \frac{\mathcal{I}(V_{\mathring{t}} + C(\Psi_k), \nu_{\mathring{t}} + 1)}{\mathcal{I}(V_0 + C(\Psi_k), \nu_0 + 1)}, \; norm = norm + w_k$$

3. Compute the probabilities

$$\alpha(K_k | d(\mathring{t})) = \frac{w_k}{norm}$$

4. Determine the final posterior pdf

$$\hat{f}(\Theta | d(\mathring{t}), K(\mathring{k})) \propto A^{\hat{\nu}_{\mathring{k}}} \exp < B(\Theta), \hat{V}_{\mathring{k}} > \chi(\Theta) \text{ with}$$

$$\hat{V}_k = \hat{V}_{k-1} + \alpha(K_k | d(\mathring{t})) C(\Psi_k), \; \hat{\nu}_k = \hat{\nu}_{k-1} + \alpha(K_k | d(\mathring{t})), \; k \in k^*$$

starting from $\hat{V}_0 = V_{\mathring{t}}, \; \hat{\nu}_0 = \nu_{\mathring{t}}$.

6 Fictitious data for normal regression model

Normal regression is the most practical instance of the parameterised model that belongs to the exponential family. It can be always reduced to the single-output case [13] we present here. The normal regression model has the form

$$m(\Psi_t, \Theta) = (2\pi r)^{-0.5} \exp\left\{\text{tr}\left[-0.5 r^{-1} [-1 \; \theta']' [-1 \; \theta'] \Psi_t \Psi_t'\right]\right\} \qquad (12)$$

where $\Theta \equiv [\theta, r] \equiv$ [vector of regression coefficients, noise variance]. A comparison of the form (12) with the general form of the exponential family (8) shows that the normal regression belongs to it with

$$A(\Theta) = (2\pi r)^{-0.5}, \; B(\Theta) = -0.5 r^{-1} [-1 \; \theta']' [-1 \; \theta'], \; C(\Psi_t) = \Psi_t \Psi_t'$$

The conjugate pre-prior pdf is a Gauss-inverse-Wishart pdf (GiW) and takes the form, cf. (9)

$$p(\Theta) \equiv GiW_\Theta(\bar{V}, \bar{\nu}) = \frac{(2\pi r)^{-0.5\bar{\nu}} \exp\left\{\text{tr}\left[-0.5 r^{-1} [-1 \; \theta']' [-1 \; \theta'] \bar{V}\right]\right\}}{\mathcal{I}(\bar{V}, \bar{\nu})}$$

with (11), see [14],

$$\begin{aligned}
V &\equiv \begin{bmatrix} V_y & V_{y\psi}' \\ V_{y\psi} & V_\psi \end{bmatrix}, \; V_y = \text{ scalar in the left upper corner of } V \\
\mathcal{I}(V, \nu) &= \left|2\pi V_\psi^{-1}\right|^{0.5} 2^{\kappa/2} \Gamma(\kappa/2) \lambda^{-\kappa/2} \qquad (13) \\
\kappa &\equiv \nu + \mathring{\theta} - 2 \\
\Gamma(x) &\equiv \int_0^\infty z^{x-1} \exp(-z) \, dz, \quad x > 0 \\
\lambda &\equiv V_y - V_{y\psi}' V_\psi^{-1} V_{y\psi} = \text{ least-squares remainder.}
\end{aligned}$$

The statistics defining GiW pdf are known to be tightly related to least squares, see e.g. [3]. They define pdf (integrable function) iff $\kappa > 0$ and V is positive definite. It is guaranteed for all data processed if these conditions are fulfilled for the statistics $\bar{V}, \bar{\nu}$ defining pre-prior pdf $p(\Theta)$. This pdf is flat for

$$\bar{V} = I\varepsilon = \text{unit matrix multiplied by a very small } \varepsilon > 0,\ \bar{\kappa} = 1.$$

These are basic facts we need for applying the general algorithm of Section 5. When applying it, we have to evaluate many times the value of $\mathcal{I}(V, \nu)$ for various V, ν. The efficient, numerically stable solution is obtained by a direct updating of a factorised version of V, [15]. It works on so-called LD factorisation of V that always exists for any positive definite V.

Proposition 6.1 *Let $V = L'DL$. L is a lower triangular matrix with a unit diagonal and D is diagonal matrix with positive entries. Then*

$$\mathcal{I}(V,\nu) = (2\pi)^{0.5\mathring{\psi}} 2^{\kappa/2}\Gamma(\kappa/2) D_{11}^{-\kappa/2} \prod_{i=2}^{\dim(\Psi)} D_{ii}^{-0.5}$$

where $\kappa = \nu + \mathring{\theta} - 2$.

Proof: Inspection of $\mathcal{I}(V,\nu)$ (13) implies that just the determinant of V_{ψ}^{-1} and the value λ have to be evaluated using the factorisation $V = L'DL$. The splitting

$$\begin{aligned} L &\equiv \begin{bmatrix} L_y & 0 \\ L_{y\psi} & L_{\psi} \end{bmatrix}, \quad L_y = 1 = \text{ scalar in the left upper corner of } L \\ D &\equiv \begin{bmatrix} D_y & 0 \\ 0 & D_{\psi} \end{bmatrix}, \quad D_y = \text{ scalar in the left upper corner of } D \end{aligned}$$

implies that

- $V_{\psi} = L'_{\psi} D_{\psi} L_{\psi} \Rightarrow |V_{\psi}| = |L_{\psi}|^2 |D_{\psi}| = 1 \cdot \prod_{i>1} D_{ii}$ and
- $\min_{\theta}[-1\ \theta'] V [-1\ \theta']' = D_{11} = D_y = \lambda.$

□

The updating of the extended information matrix V has the structure $V_{new} = V + \Psi\Psi' \Leftrightarrow L'_{new} D_{new} L_{new} = L'DL + \Psi\Psi'$. The direct re-computation L, D, Ψ to L_{new}, D_{new} can be performed efficiently using, for instance, the algorithm dydr described in [16].

6.1 Scaling of information pieces for normal regression

Usual variants of the prior information

outputs $y \in (\underline{y}, \overline{y}) \equiv$ are highly expected for a given regression vector ψ

are discussed in subsequent Section. We would like to interprete knowledge of the (soft) bounds $\underline{y} < \overline{y}$ (for the given ψ) as fictitious measurements. It is reasonable to take middle of the interval $\hat{y} = (\overline{y} + \underline{y})/2$ as the expected value of y and the half-width of this interval as its standard deviation $\sqrt{c_y} = (\overline{y} - \underline{y})/2$.

We search for a fictitious data vector Ψ_K reflecting "just" the information K. It means that the restrictions on the constructed prior pdf $f(\Theta|K) \equiv f(\Theta|\Psi_K)$

$$\begin{aligned} \hat{y} &= \int y f(y|\psi, K)\, dy \quad c_y = \int (y - \hat{y})^2 f(y|\psi, K)\, dy \text{ with} \qquad (14) \\ f(y|\psi, K) &= \int f(y|\psi, \Theta) f(\Theta|\Psi_K)\, d\Theta \end{aligned}$$

are fulfilled and $f(\Theta|\Psi_K) \propto m(\Psi_K, \Theta)p(\Theta)$ is as close as possible to $p(\Theta)$. Again, we use Kullback-Leibler distance as the distance measure between considered pdfs.

The unknown parameters Θ consist of regression coefficients θ and the noise variance r. We expect that the output variance c_y is greater than the noise variance r, i.e. $c_y > r \Rightarrow c_y > \bar{r} \equiv \int r p(r)\, dr$ & $c_y > \int r f(r|K)\, dr$. Here, we assume that the inequality

$$c_y > \bar{r} \Leftrightarrow \zeta_y^{-1} \equiv c_y/\bar{r} - 1 > 0 \qquad (15)$$

is fulfilled. Then, we can require $f(r|K) = p(r)$, i.e. to adopt assumption that the processed information piece brings nothing new about r.

It can be shown that the statistics V, ν determining the constructed GiW pdf should have the form

$$\begin{aligned} V_\psi &= \bar{V}_\psi + b\psi\psi' \qquad (16) \\ V_{y\psi} &= \bar{V}_{y\psi} + a\psi \\ \nu &= \bar{\nu} \\ \lambda &= \bar{\lambda} \equiv \bar{V}_y - \bar{\theta}'\bar{V}_\psi \bar{V}_y \bar{\theta}, \text{ with } \bar{\theta} \equiv \bar{V}_\psi^{-1}\bar{V}_{y\psi} \\ &\Leftrightarrow \ V_y = \bar{\lambda} + (\bar{V}_{y\psi} + a\psi)'(\bar{V}_\psi + b\psi\psi')^{-1}(\bar{V}_{y\psi} + a\psi). \end{aligned}$$

The optional scalars a, b have to be chosen so that restrictions (14) are met, i.e.

$$\hat{y} = \hat{\theta}'\psi, \text{ with } \hat{\theta} = V_\psi^{-1}V_{y\psi} \quad c_y = \bar{r}(1 + \psi' V_\psi^{-1}\psi) \equiv \frac{\bar{\lambda}}{\bar{\nu} - 2}(1 + \psi' V_\psi^{-1}\psi). \qquad (17)$$

The constructed function should be pdf, i.e. V_ψ has to be positive definite. Thus, the number $\zeta \equiv \psi' V_\psi^{-1}\psi$ has to be necessarily positive. The construction below shows that this condition is also sufficient.

Under the adopted assumption (15), the second restriction in (17) gives (with some algebra)

$$b = \zeta_y^{-1} - \bar{\zeta}^{-1}, \text{ with } \bar{\zeta} = \psi'\bar{V}_\psi\psi. \qquad (18)$$

Note that this weight might be negative. However, expressing the matrix $V_\psi = \bar{V}_\psi^{0.5}\left[I + b\bar{V}_\psi^{-0.5}\Psi\left(\bar{V}_\psi^{-0.5}\Psi\right)'\right](\bar{V}_\psi^{0.5})'$, we see that it is positive definite if the expression in brackets [ ] is positive. The term in brackets is symmetric matrix. It has $\mathring{\psi}$ eigenvalues equal to 1. The remaining eigenvalue equals to $1 + b\bar{\zeta} = \bar{\zeta}/\zeta_y > 0$. Thus, the matrix V_ψ is positive definite iff (15) is fulfilled.

The first restriction in (17) provides the equation for a that gives

$$a = \frac{\hat{y}}{\zeta_y} - \frac{\bar{y}}{\bar{\zeta}}, \text{ with } \bar{y} = \bar{\theta}'\psi = \bar{V}'_{y\psi}\bar{V}_\psi^{-1}\psi.$$

The third restriction in (16) is simple, thus, it remains to fulfil the fourth one in (16). Inserting, a, b into it, we get

$$V_y \equiv \bar{V}_y + \frac{\hat{y}^2}{\zeta_y} - \frac{\bar{y}^2}{\bar{\zeta}}. \tag{19}$$

As V_ψ and $\lambda \equiv \bar{\lambda}$ are positive definite by construction, the constructed statistics V will be positive definite iff the scalar V_y in (19) will be positive. Generally, it has to be checked. There is, however, a special but practically dominating case with the pre-prior pdf expecting zero regression coefficients

$$\bar{\theta} = 0 \Rightarrow \bar{y} = 0. \tag{20}$$

When considering it, the positiveness of V_y is guaranteed automatically and *the inequality (15) is the only condition of the solvability of the formulated task.*

Proposition 6.2 *Let us consider estimation of parameters of normal regression model. Let the pre-prior pdf fulfil (20) and the related expected value of the noise variance r be $\bar{r} = \varepsilon$.*

The fictitious data reflecting prior information about

- *the expected value $\hat{y}$ and variance c_y of the output of the normal regression conditioned on a given regression vector ψ,*
- *the prior distribution of noise variance $f(r|K) = p(r)$*

exist iff $\zeta_y \equiv c_y/\bar{r} - 1 > 0$.

If, moreover, $\bar{\zeta}^{-1} \equiv (\psi'\bar{V}_\psi^{-1}\psi)^{-1} = 0$ then the corresponding fictitious data are simply

$$\Psi'_K := \frac{1}{\sqrt{\zeta_y}}[\hat{y}, \psi]. \tag{21}$$

Proof: The majority of statements of this proposition have been demonstrated above. It remains to notice that $b > 0$ is always positive if $\bar{\zeta}^{-1} = 0$, see (18). Moreover, the weight in (21) guarantees that the moments implied by (16), (17) are attained. □

Note, that the improper pre-prior pdf leads to the intuitively appealing scaling (21). In this case, $\bar{V}$ contains just single non-zero element $\bar{V}_y = \varepsilon$. It is acceptable whenever the real observed data create the proper posterior pdf, i.e. when the matrix $V_{\mathring{t}} \equiv \sum_{t \in t^*} \Psi_t \Psi'_t$ is regular.

The possibility of use of the improper pre-prior pdf $p(\Theta)$ has to be carefully checked. For instance, the evaluation of the weights $\alpha(K_k|d(\mathring{t}))$ uses also this prior. If it is updated by a single data vector only the proper pdf is not obtain as a rule. A similar care is needed in structure estimation [17] tasks.

6.2 Examples of information pieces for normal regression

In order to outline applicability of the approach, it is worthwhile to list how some pieces of information can be translated into fictitious data. In all cases, linear-in-data regression model is used with $\psi'_t = [y_{t-1}, \ldots, y_{t-m_y}; u_t, \ldots, u_{t-m_u}]$ for some finite $m_y, m_u \geq 0$ and scalar outputs y and inputs u. We also assume that uncertainty c_y about any information piece is realistic with respect to the pre-prior estimate $\bar{r} = \varepsilon > 0$ of the noise covariance, i.e. $\zeta_y = c_y/\varepsilon - 1 > 0$.

Static gain g: for unit inputs, the output stabilises around the value g.

Let we expect $g \in (\underline{g}, \overline{g})$. Then $\hat{y} = 0.5(\overline{g} + \underline{g})$, $\sqrt{c_y} = 0.5(\overline{g} - \underline{g})$.
Thus, the scaled fictitious data are

$$\Psi = \frac{1}{\sqrt{\zeta_y}}[\hat{y}; \psi'] = \frac{1}{\sqrt{\zeta_y}}[\hat{y}; \hat{y}, \ldots, \hat{y}; 1, \ldots, 1].$$

Dominating time constant T: output response to non-zero initial conditions diminishes approximately as a single exponential.

Let T_s be a sampling period of the continuous-time system and dominating time constant T is in the range $(\underline{T}, \overline{T})$. Then, for $y_0 = 1$, $y_\tau = 0, \tau < 0$ and $u_t = 0$, $\forall t$, we expect $y_t \in (\underline{y}_t, \overline{y}_t) \equiv \left(\exp\left[-\frac{tT_s}{\underline{T}}\right], \exp\left[-\frac{tT_s}{\overline{T}}\right]\right)$

These data may be generated by the first order difference equations
$\underline{y}_t = \underline{A} y_{t-1}$, $\overline{y}_t = \overline{A} y_{t-1}$ with $y_0 = 1$, $\underline{A} = \exp\left[-\frac{T_s}{\underline{T}}\right]$, $\overline{A} = \exp\left[-\frac{T_s}{\overline{T}}\right]$.

Defining $\hat{y}_t = 0.5(\overline{y}_t + \underline{y}_t)$, for $t \geq 1$ and $\hat{y}_0 = 1$, $\hat{y}_\tau = 0$, $\forall \tau < 0$ and $\sqrt{c_{y_t}} = 0.5(\overline{y}_t - \underline{y}_t)$, we get the scaled fictitious data

$$\Psi_t = \frac{1}{\sqrt{\zeta_{y_t}}}[\hat{y}_t; \psi'_t] = \frac{1}{\sqrt{\zeta_{y_t}}}[\hat{y}_t; \hat{y}_{t-1}, \ldots, \hat{y}_{t-m_y}; 0, \ldots, 0].$$

This sequence is always finite as it is restricted by the requirement $\zeta_{y_t} \equiv c_{y_t}/\bar{r} - 1 > 0$.

A block treatment, see Section 7, may be useful for this piece of information. The extension to more dominant time-constants is straightforward. Practically, it makes sense to specify a pair of such constants. The characterisation of time constants as a reaction on initial conditions allows us to insert this information independently of the gain of the system.

Step response: for unit step of inputs, the output runs around step response.

The treatment follows general theory with individually handled fictitious outputs and regression vectors. The restriction $\zeta_{y_t} > 0$ (15) serves as the natural stopping rule for the number of data to be processed.

Frequency response: for inputs of the form $\sin(\omega t)$ with a fixed frequency ω, the outputs get the form $A(\omega)\sin(\omega t + \phi(\omega))$. The pair $A(\omega), \phi(\omega)$

determines a point on frequency response. $A(\omega)$ is called amplitude and $\phi(\omega)$ is phase (at the frequency ω).

The range of the possible outputs is given through a pair of information pieces, namely, $A(\omega) \in (\underline{A}(\omega), \overline{A}(\omega)$ and $\phi \in (\underline{\phi}(\omega), \overline{\phi}(\omega))$. The second one, phase, influences this range in a non-linear fashion through the goniometric function, which has the range [-1,1]. Consequently, we can safely assume that the expected range of the output and the range of $A(\omega)$ coincide. Thus, for any $\phi(\omega)$ we assume $y \in (\underline{A}(\omega), \overline{A}(\omega))$. It leads us to the choice of the expected output $\hat{y} = 0.5(\underline{A}(\omega) + \overline{A}(\omega))$ and $\sqrt{c_y} = 0.5(\overline{A}(\omega) - \underline{A}(\omega))$.

For a fixed $\phi(\omega)$, the possible fictitious data vectors are of the form

$$[\hat{y}\sin(\omega t + \phi(\omega)); \hat{y}\sin(\omega(t-1) + \phi(\omega)), \ldots, \hat{y}\sin(\omega(t-m_y) + \phi(\omega)); \\ \sin(\omega t), \ldots, \sin(\omega(t-m_u)].$$

Any of them can be written as a linear combination of the pair of fixed vectors

$$\begin{aligned}
\Psi_1 &\equiv [\hat{y}\sin(\phi(\omega)); \hat{y}\sin(-\omega + \phi(\omega)), \ldots, \hat{y}\sin(-\omega m_y + \phi(\omega)); \\
&\quad \sin(0), \ldots, \sin(-\omega m_u)] \\
\Psi_2 &\equiv [\hat{y}\cos(\phi(\omega)); \hat{y}\cos(-\omega + \phi(\omega)), \ldots, \hat{y}\cos(-\omega m_y + \phi(\omega)); \\
&\quad \sin(0), \ldots, \cos(-\omega m_u)]
\end{aligned}$$

that serve as (non-normalised) fictitious data vectors expressing the information on frequency response (for a fixed $\phi(\omega)$).

In order to cover uncertainty both in amplitude and phase, it is necessary to select several pairs of such data vectors differing in phase. At present, two pairs corresponding to boundaries $\underline{\phi}(\omega)$, $\overline{\phi}(\omega)$ are successfully used.

Scaling of the discussed fictitious data vectors is done in the standard common way.

Note that ω has to be expressed in discrete time considered. It means that $\omega = T_s\omega_c \equiv$ sampling period $\times$ continuous frequency.

Cut-off frequency ω_c: for periodic inputs with a frequency above the cut-off frequency, the output cannot be distinguished from the noise. Thus, we have a special information on frequency response with $A(\omega) \in c(-\sqrt{\bar{r}}, \sqrt{\bar{r}})$ with $c > 1$ reflecting our confidence into this information. As the amplitude is supposed to be (practically) zero for all frequencies above ω_c we process several fictitious data vectors for $\omega = l\omega_c$, $l \in l^* \equiv \{1, \ldots, \mathring{l} < \infty\}$

Note that the conversion of continuous-time information has to be converted into the sampled world as in the general case.

Dynamic data consist either from operational input-output records that are suspicious for being measured at different working conditions or from records from a (simulated) model of the system modelled.

These data can be directly used as fictitious ones if we are able to determine their uncertainty. It can be done by performing parameter estimation solely on them. It provides, among others, the estimate of the output variance c_y that is used in the general scaling.

Sparse data are input-output pairs recorded in an uneven way, for instance from graphs of some static characteristics. In this case, a sort of a simple interpolation is used for filling fictitious data vector(s). Otherwise, they are treated as dynamic data.

The other possible information pieces like smoothness of step response, time constant for multiple poles can be prepared in similar way.

7 Extension to blocks of fictitious data

Obsolete and simulation data are a very important source of prior information. They can be treated using the described methodology if their cardinality is not very large. The exploitation of large data sets needs an efficient solution of processing of data blocks. Data within an internally consistent block can be combined with a constant (geometric) weight of individual items and then block combined in a way mimic to the approach described above. We discuss this technique here.

Let us consider a piece of information K_k expressed by finite sequence of fictitious of "compatible" fictitious data vectors $\Psi_{i_k k},\ i_k \in i_k^* \equiv \{1, \ldots, \mathring{i}_k\}$. Compatibility means that the $\alpha(K_{i_k k}) \equiv$

$$\text{the prior belief that } \left[f(\Theta|K_{i_k k}),\ i_k \in i_k^* \ \text{ is closest to } f(\Theta|K(\mathring{k})) \right] = \frac{1}{\mathring{i}_k}.$$

Then, the pdf

$$\hat{f}(\Theta|K_k) \propto \prod_{i_k \in i_k^*} [m(\Psi_{i_k k}, \Theta)]^{1/\mathring{i}_k}\, p(\Theta) \equiv M(k, \Theta) p(\Theta)$$

has the smallest expected Kullback-Leibler distance to the unknown $f(\Theta|K_k(\mathring{i}_k))$. The model $M(k, \Theta)$ correcting the pre-prior pdf $p(\Theta)$ has the form

$$M(k, \Theta) \equiv \prod_{i_k \in i_k^*} [m(\Psi_{i_k k}, \Theta)]^{1/\mathring{i}_k}.$$

The rest of the computation is similar to that presented in Section 3, i.e.

$$\hat{f}(\Theta|d(\mathring{t}), K(\mathring{k})) \propto L(\Theta, d(\mathring{t})) \prod_{k \in k^*} [\tilde{M}(k, \Theta)]^{\alpha(K_k|d(\mathring{t}))}$$

with $\tilde{M}(k, \Theta) = m(\Psi_k, \Theta)$ for individual data vectors and $\tilde{M}(k, \Theta) = M(k, \Theta)$ for data blocks. The likelihood $L(\Theta, d(\mathring{t}))$ is defined by the formula (6) and probabilities $\alpha(K_k|d(\mathring{t}))$ assigned to alternative information pieces are

$$\alpha(K_k|d(\mathring{t})) \propto \frac{\int \mathcal{L}(\Theta, d(\mathring{t}))p(\Theta)\tilde{M}(k,\Theta)\,d\Theta}{\int p(\Theta)\tilde{M}(k,\Theta)\,d\Theta}.$$

For exponential family, see Section 4, we get the following straightforward extension of the algorithm presented in Section 5

Preparatory phase

1. Select parameterised model (its structure).
2. Specify pre-prior information by selecting statistics $\bar{V}$, $\bar{\nu}$ and support of the constructed pdf.
3. Compute statistics corresponding to the posterior likelihood function assigned to the pre-prior pdf, i.e.

   $$V_t = V_{t-1} + C(\Psi_t),\ \nu_t = \nu_{t-1} + 1,\ t \in t^*,\ V_0 = \bar{V},\ \nu_0 = \bar{\nu}.$$

   The final likelihood function corresponding to the vague pre-prior pdf $p(\Theta)$ has the form

   $$L(\Theta, d(\mathring{t})) \equiv L(\Theta, V_{\mathring{t}}, \nu_{\mathring{t}}) = A^{\nu_{\mathring{t}}}(\Theta) \exp < B(\Theta), V_{\mathring{t}} > \chi(\Theta).$$

Processing of information items

1. Set the auxiliary variable $norm = 0$ and
2. Do, for $k \in k^*$, while information pieces are available:
   (a) Express the information item K_k through fictitious data Ψ_k if single fictitious data vector is processed ($\mathring{i}_k = 1$, i.e. set $\tilde{C}_k \equiv C(\Psi_k)$ or Express increment of the sufficient statistics

   $$\tilde{C}_k = \frac{1}{\mathring{i}_k} \sum_{i_k \in i_k^*} C(\Psi_{i_k k})$$

   where $\Psi_{i_k k}$ are individual fictitious data vectors expressing bits of the kth piece of information.
   (b) Scale $\tilde{C}_k$ (see below).
   (c) Compute the weight w_k and add it to $norm$.

   $$w_k = \frac{\mathcal{I}(V_{\mathring{t}} + \tilde{C}_k, \nu_{\mathring{t}} + 1)}{\mathcal{I}(V_0 + \tilde{C}_k, \nu_0 + 1)},\ norm = norm + w_k$$

3. Compute the probabilities

   $$\alpha(K_k|d(\mathring{t})) = \frac{w_k}{norm}$$

4. Determine the final posterior pdf

$$\hat{f}(\Theta|d(\mathring{t}), K(\mathring{k})) \propto A^{\hat{\nu}_{\mathring{k}}} \exp < B(\Theta), \hat{V}_{\mathring{k}} > \chi(\Theta) \text{ with}$$

$$\hat{V}_k = \hat{V}_{k-1} + \alpha(K_k|d(\mathring{t}))\tilde{C}_k, \ \hat{\nu}_k = \hat{\nu}_{k-1} + \alpha(K_k|d(\mathring{t})), \ k \in k^*$$

starting from $\hat{V}_0 = V_{\mathring{t}}, \ \hat{\nu}_0 = \nu_{\mathring{t}}$.

It remains to discuss how to scale increments of sufficient statistics $\tilde{C}_k$, see step 2b. We consider data blocks ($\mathring{i}_k > 1$) as the case $\mathring{i}_k = 1$ is fully described and specialised to normal models above. The restrictions (14) are fulfilled if we scale the individual data vectors $\Psi_{i_k k}$ as in the case $\mathring{i}_k = 1$. Then we have to care about the restriction $f(r) = p(r)$ only. Generally, it cannot be met without violating (14). The problem can be addressed in the following approximate way. The constructed $\tilde{C}_k$, added to $\bar{V}$, defines pdf $f(\Theta) \equiv f(\theta, r) = f(\theta|r)f(r)$ that respects the restrictions (14) but violates the requirement $f(r) = p(r)$. The pdf $\hat{f}(\Theta) = f(\theta|r)p(r)$, that fulfils it, can be shown as the pdf closest (in terms of Kullback-Leibler distance) to $f(Theta)$. Thus, it is reasonable to use it as the final solution.

It is straightforward to find how this solution looks in the case of normal regression. Let V be the statistics obtained for the considered data block with the LD factorisation used in the proof of Proposition 6.1. Then the restriction $f(r) = p(r)$ is met simply by replacing D_y with $\bar{r} = \varepsilon$.

8 Implementation in ABET

The problem addressed within the paper has been motivated by a project DESIGNER [18, 6] that aims at developing a complete algorithmic support for prior tuning of adaptive controllers. At present, a full version is available for single-input single-output systems modelled by linear regression model with Gaussian disturbances and quadratic performance index (LQG). Majority of building blocks is ready for multivariate case. The DESIGNER is based on general MATLAB functions creating core of *Adaptive Bayesian Estimation Toolbox* (ABET) [19, 20] covering key steps of the design:

- data preprocessing
- processing of prior information
- tuning of forgetting
- control loop performance prediction
- structure estimation
- parameter estimation
- control design
- control verification

The individual ABET functions might be of independent interest. Here, we are mentioning them for describing environment into which the proposed theory is implemented. Moreover, illustrative examples presented in the next section were generated by ABET.

9 Illustrative experiments

Improved pre-tuning of adaptive controllers has been primary motivation for the research described. Thus, it is natural that we illustrate the influence of the prior information on the control quality achieved by a controller that uses it. Results of structure and forgetting-factor estimation are also presented as additional indicators of the inspected influence. Especially forgetting factor has been found as a sensitive indicator of the model reliability: the closer its estimate is to unity the more confident we can be about model suitability.

9.1 Simulated system

The simulated continuous-time system has the following transfer function

$$F(s) = \frac{-1.0882s^3 + 81.473s^2 - 3378s + 58647}{2.2388s^3 + 941.88s^2 + 598.97s + 54517}. \tag{22}$$

It was sampled with the sampling period $T_s = 0.05$ and converted into the difference equation

$$\begin{aligned} y_t &= \theta\psi' + e_t = \\ &= [1.9919,\, -2.2026,\, 1.8408,\, -0.8941,\, 0.1028,\, 0.1812] \times \\ &\times\ [y_{t-1},\, y_{t-2},\, y_{t-3},\, y_{t-4},\, u_{t-3},\, u_{t-4}]' + e_t \end{aligned} \tag{23}$$

to which a normal zero mean white noise e_t was added. Noise variance $r = 0.002$ was chosen.

Data used for pre-tuning of linear-quadratic-Gaussian (LQG) adaptive controller were generated by SIMULINK scheme shown in Figure 1. The block, called *Discrete Transfer Function 1*, is our system itself. The noise e_t is generated by the block *Band-Limited White Noise.* Moreover, measurement noise *Band-Limited White Noise 2* with variance 0.0006 is added in order to make experiments close to conditions met in practice.

9.2 Experimental set-up

Pre-tuning of the LQG controller with ABET was done according to the user wish to minimise output variance around output reference $y_{ref} = 0$ while keeping inputs inside the interval [-1,1]. This restriction determines penalisation of inputs.

Preliminary data of lengths 100, 200, 500 and 1000 measured on the system stimulated in open loop by white noise were used. Pre-tuning was used with no prior information and prior information on the static gain $g \in [1, 1.1]$.

Quality of pre-tuning was verified by closed loop experiments with LQG adaptive controller. The verification run always with $\mathring{t} = 1000$.

The sample standard deviation of the output

$$v_y = \sqrt{\frac{\sum_{t=1}^{\mathring{t}} (y_t - y_{ref})^2}{\mathring{t}}},$$

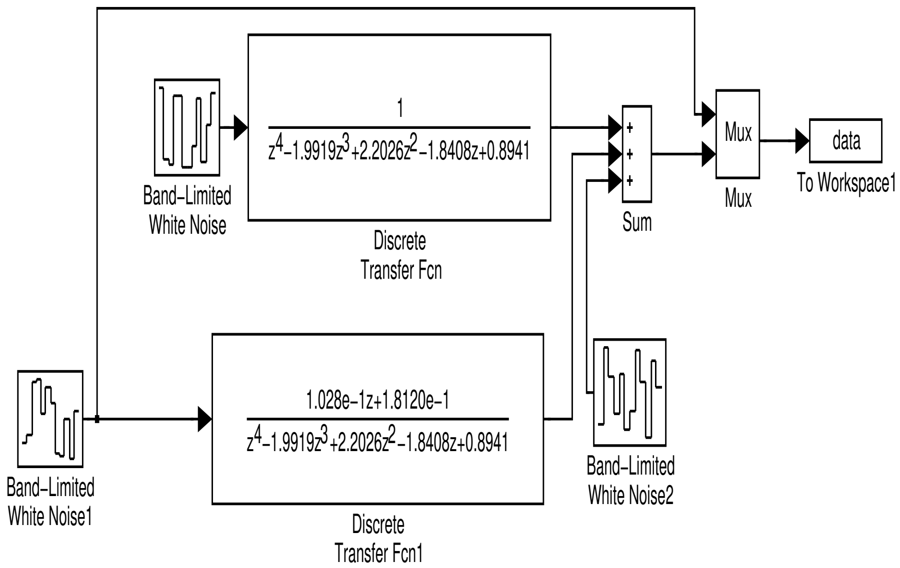

Figure 1: **Test system – open loop scheme**

was used as a quality indicator. Moreover, the practically important range of inputs is recorded as an indicator of the design success (the requirement on it is translated into input penalisation in control design block of ABET).

9.3 Verification results

The results of pre-tuning obtained for 100 preliminary data, demonstrate well influence of the prior information.

In both cases, with and without prior information, the results of structure and parameter point estimate coincide

$$\begin{aligned} \hat{y}_t &= [1.1569,\, -0.5386,\, 0.6444,\, -0.7471,\, 0.0472,\, 0.1223,\, 0.0694] \times \\ &\times\ [y_{t-1},\, y_{t-2},\, y_{t-4},\, y_{t-5}\, u_{t-3}\, u_{t-4}\, u_{t-5}]'. \end{aligned}$$

Point estimates cannot be compared with the parameters of the simulated system as the model structure is over-estimated. The achieved control quality and recommended forgetting factors (indicating model reliability) differ substantially as seen in Table 1.

Table 1: **Influence of prior information on model reliability and control quality – 100 learning data**

	forgetting factor	input penalty	range of inputs	standard deviation of outputs v_y
no prior	0.8615	1.41e-4	[-2.357 2.457]	1.074
with prior	0.9999	1.41e-4	[-0.757 0.766]	0.061

The positive influence of the prior information is obvious. It allowed us to improve significantly forgetting factor, input ranges and output variance. It is, however, fair to warn against over-optimism. Sometimes no improvements or even a slight deterioration may take place. The cases of this type are in Table 2 that shows indicators of the closed quality in dependence on the length of data used in pre-tuning.

Table 2: **Influence of prior information model reliability and control quality – dependence on learning data-length**

learning-data length	200	200	500	500	1000	1000
	no prior	prior	no prior	prior	no prior	prior
forgetting factor	0.9398	0.9999	0.9878	9.9999	0.9958	0.9958
input range	[-1.314 1.298]	[-0.837 0.833]	[-0.885 0.891]	[-0.839 0.833]	[-0.785 0.788]	[-0.803 0.801]
standard deviation v_y	0.112	0.057	0.062	0.057	0.046	0.054

For 1000 learning data items, prior information made the closed loop behaviour a little bit worse than without its use. It might be attributed either to its approximate expression or random effects. In any case, the role of prior information decreases with increasing of data length as it could be expected.

Many other experiments have been performed. They confirmed the qualitatively expected behaviour: significant information pieces get higher weights $\alpha(K_k|d(\mathring{t}))$ and, importantly, those contradicting real data have these weights negligible. Consequently, the control quality can hardly be completely spoiled by using incorrect or superfluous information.

10 Conclusions

Universality and simplicity of black-box based adaptive controllers are the main attributes that determine their high practical potential. It is, however, inhibited by a high number of optional parameters to be tuned. This is a hard task for non-experts. Improper choices lead to wild transients and lower-than-possible quality. This motivated a long term project DESIGNER that aims at automatic pre-tuning of adaptive controllers of this type using user's wishes, restrictions, prior data and expert information.

This paper deal with the last item, namely, quantitative expression of prior information contained both in prior data and expert information. The solution suits to processing global widely available information like information on static gain, frequency response, artificial data obtained from a complex system model etc. It is conservative in a way that is robust with respect to repetitive information items and items that contradicts data measured. The solution is implemented in software toolbox ABET and improves properties of the tuned adaptive controller as expected. The toolbox ABET is available via anonymous ftp at ftp.utia.cas.cz in the directory /pub/staff/nedoma.

The presented solution could be used out of the motivating area for improving other adaptive systems like estimators, predictors, fault-detectors etc. The proposed solution is built under a definite information scenario (parameterised priors are tuned) so that it should be taken as a complement of other well-established techniques of prior information elicitation.

Acknowledgment This research has been partially supported by GAČR No. 102/99/1564, No. 102/99/1292 and EU project IST-1999-12058.

References

[1] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, London, 1987.

[2] T. Bohlin, *Interactive System Identification: Prospects and Pitfalls*, Springer-Verlag, Berlin, Heidelberg, New York, 1991.

[3] V. Peterka, “Bayesian system identification”, in *Trends and Progress in System Identification*, P. Eykhoff, Ed., pp. 239–304. Pergamon Press, Oxford, 1981.

[4] J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York, 1985.

[5] M. Kárný and P. Nedoma, “Automatic processing of prior information with application to identification of regression model”, *Kybernetika*, 2000, submitted.

[6] J. Bůcha, M. Kárný, P. Nedoma, J. Böhm, and J. Rojíček, “Designer 2000 project”, in *International Conference on Control '98*, London, September 1998, pp. 1450–1455, IEE.

[7] P. Nedoma, M. Kárný, and J. Böhm, “ABET99. Adaptive Bayesian Estimation Toolbox for MATLAB”, Tech. Rep. 1963, ÚTIA AV ČR, Praha, 1999.

[8] M. Kárný, J. Šindelář, and M. Valečková, "Towards Bayesian pooling of probabilistic controllers", in *Preprints of the 3rd European IEEE Workshop on Computer-Intensive Methods in Control and Data Processing*, J. Rojíček, M. Valečková, M. Kárný, and K. Warwick, Eds., Praha, September 1998, pp. 19–24, ÚTIA AV ČR.

[9] M. Kárný, M. Valečková, and H. Gao, "Mixtures of adaptive controllers based on Markov chains: A future of intelligent control?", in *International Conference on Control '98*, London, September 1998, pp. 721–726, IEE.

[10] S. Kullback and R. Leibler, "On information and sufficiency", *Annals of Mathematical Statistics*, vol. 22, pp. 79–87, 1951.

[11] L. Berec and M. Kárný, "Identification of reality in Bayesian context", in *Computer-Intensive Methods in Control and Signal Processing: Curse of Dimensionality*, K. Warwick and M. Kárný, Eds. Birkhauser, 1997.

[12] O. Barndorff-Nielsen, *Information and exponential families in statistical theory*, Wiley, New York, 1978.

[13] M. Kárný, "Parametrization of multi-output multi-input autoregressive-regressive models for self-tuning control", *Kybernetika*, vol. 28, no. 5, pp. 402–412, 1992.

[14] M. Kárný, "Identification for adaptive control", Tech. Rep., UTIA AV ČR, POBox 18, 182 08 Prague 8, Czech Republic, 1990, DrSc thesis, in Czech.

[15] G.J. Bierman, *Factorization Methods for Discrete Sequential Estimation*, Academic Press, New York, 1977.

[16] V. Peterka, "Control of uncertain processes: applied theory and algorithms", *Kybernetika*, vol. 22, 1986, Supplement to No. 3, 4 ,5, 6.

[17] M. Kárný and R. Kulhavý, "Structure determination of regression-type models for adaptive prediction and control", in *Bayesian Analysis of Time Series and Dynamic Models*, J.C. Spall, Ed. Marcel Dekker, New York, 1988, chapter 12.

[18] P. Nedoma, and M. Kárný and J. Böhm, "Project DESIGNER", in *Adaptive and Predictive Control, Proceedings of the Spring school*, A. Kuznetsov, Ed., pp. 91–93. Oxford University, UK, 1996.

[19] P. Nedoma, M. Kárný, and J. Böhm, "Software tools for use of prior knowledge in design of LQG adaptive controllers", in *Preprints of the IFAC Workshop on Adaptive Systems in Control and Signal Processing*, Glasgow, August 1998, pp. 425–429, IFAC.

[20] P. Nedoma, M. Kárný, J. Böhm, J. Rojíček, and L. Berec, "ABET98. Adaptive Bayesian Estimation Toolbox for MATLAB", Tech. Rep. 1937, ÚTIA AV ČR, Praha, 1998.

file=acaspri.tex, revised version of prirev1.tex, April 6, 2000